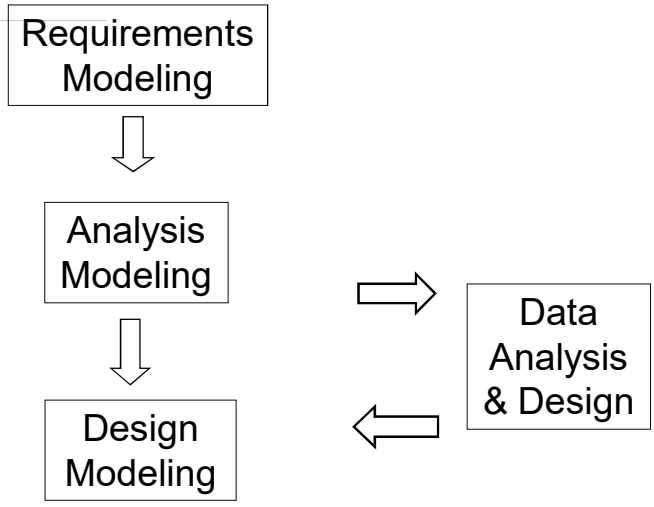
Data Analysis and Design

- Data Analysis

Instructor Chia Yuen Kwan
Email: isscyk@nus.edu.sg
Institute of Systems Science, National University of Singapore
25 Heng Mui Keng Terrace, Singapore 119615

Application Development Workflow

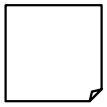


In this Lecture you will learn

- About
  - Data Analysis
    - Entity Relationship diagram (data model)
    - Attribute Analysis
  - Data Design
    - Normalization
    - Normalized Data Model

Why Data Analysis and Design?

- Develop a data model for data required by the business requirements
  - A data model consists of:
    - Entities Relations Diagram
    - Data Dictionary

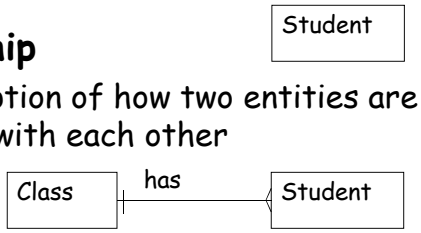


- Required before designing and implementing database

Entity Relationship Diagram or ERD

Entity Relationship Diagram

- Is a graphic representation of how structured data are being organized
- Components include:
  - A business **Entity**
    - An object about which information is to be maintained
  - A **Relationship**
    - Is a description of how two entities are associated with each other

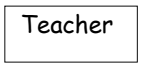


Entities

- A distinguishable objects in the problem domain that we want to model.
- You need to distinguish:
  - Entity Type (or Entity)
  - Entity Occurrence

In a ISS Course Registration System

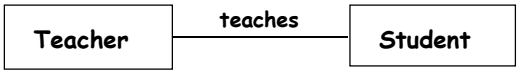
- Teacher is an Entity Type



- Esther is the Entity Occurrence

Relationships

- Relations between entity types



- Cardinality of Relationship
  - The number of occurrences of one entity type that relate to the occurrences of another entity type

    one- to- one

    one- to- many

    many-to-many

    etc

Notations: Cardinality of Relationships

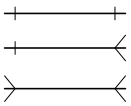
- Cardinality of Relationship
  - The number of occurrences of one entity type that relate to the occurrences of another entity type

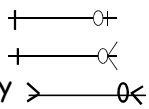
    one- to- one

    one- to- many

    many-to-many

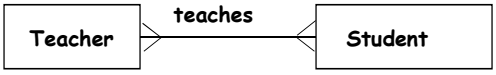
    etc

    
  - Optional Relationships
    - one to zero or one
    - one to zero or many
    - many to zero or many
    - etc

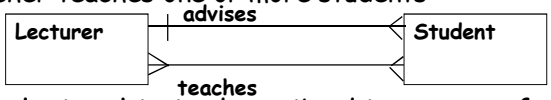
    

Relationships

- Properties of Relationship
  - Depends on the business rules
  - Every Relationship is bi-directional
    - a teacher teaches one or more students
    - a student is taught by one or more teacher

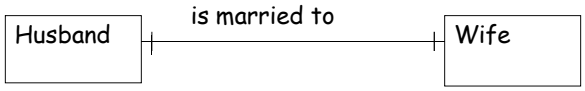
  - There may be more than one <u>important relationships</u>
    - a teacher counsels zero, one or more students
    - a teacher teaches one or more students

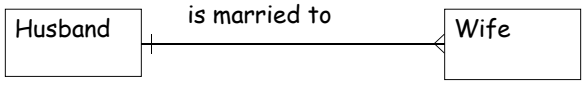
  - Every Relationship is described in terms of a verb

Example

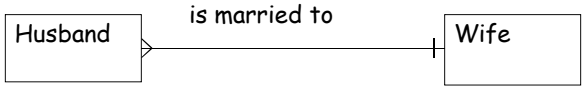
Monogamy society



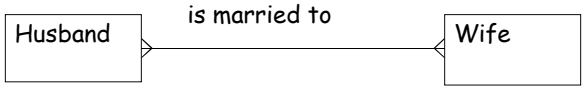
Muslim world



Africa tribe

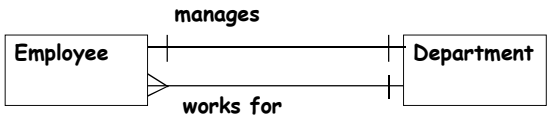


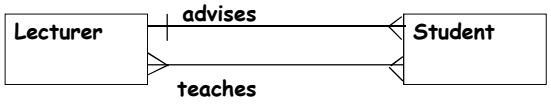
A religion Cult



Example

- More than one <u>important</u> relationships





Example: Project Management System

- System is to keep track of
  - all the projects undertaken by the company
  - assignments of projects to departments
    - a project can be assigned to one or more departments
    - each department can take on one or more projects
    - one or more employee of a department may work on a project
    - a employee can only take on one project at any one time
    - a non-active project (e.g. a cancelled project) may not be assigned to any employee or under any department

Project Management System

How would the ERD looks like?

Project Management Systems

- Possible attributes of each entity are
  - Employee
    - employee name,job title, job description, project name, employee skill type, employee skill type description, employee monthly salary etc
  - Department
    - department name, department manager number, department manager name, department employee size, project name, project-department budget allocated, department employee number, department employee name
  - Project
    - project name, project description, project budget allocated, project start date, project end date

Attributes Analysis

Attribute Analysis

- During data analysis, a number of attributes would have been discovered.
- Adopt and/or establish <u>good attribute(data) naming standard</u>

Data Naming Standard

- Attribute Names should be
  - FULL/COMPLETE
  - MEANINGFUL

Example:

Hand-Phone-Number	vs	**Phone-Number**
Passport-Number	vs.	**Number**

Data Naming Standard (cont'd)

- Guideline on constructing an attribute name
  - Identify keywords from the description of the attribute
  - Arrange keywords in certain sequence
  - Join the keywords with hyphen

Example:

Hand-Phone-Number

Passport-Number

Data Naming Standard -Guidelines

CLASS KEYWORDS

- Classification of the type of data item values or data types

PRIME KEYWORDS

- Describe the prime business object or process to which the data item pertains

MODIFIER KEYWORDS

- A keyword that "modifies" the prime keyword
- If more than one modifier keywords are needed, they should be ordered from left to right, from most specific to most general
- For time-dependent modifier keyword, it should be positioned immediately before the keyword it modifies

EXAMPLE

An attribute "Salary" has this description
"Salary paid to the employee from the beginning of the current year to present"

Class Keyword	→	Amount
Prime Keyword	→	Salary
Modifier Keyword	→	Employee
Time-dependent Modifier Keyword	→	Year To Date

some programming languages come with their own naming guidelines too

Resulting Attribute Name

Employee-Year-To-Date-Salary-Amount

Data Naming Standard

If names are full, complete, etc.....

They may NOT be IMPLEMENTABLE

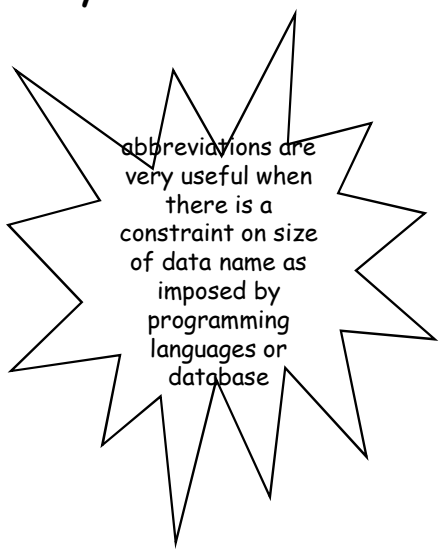
in the target language, database

To Overcome This Problem....

- Replace the name with abbreviations
- Enter the results in a DATA DICTIONARY and provide aliases in the dictionary

Class Keyword Glossary

Keyword	Abbreviation
Amount	A
Name	M
Code	C
Number	N
Constant	K
Percent	P
Date	D
Quantity	Q
Flag	F
Switch	C
Time	H
Description	DESC
Text	T



Keyword Glossary

Keyword	Abbreviation
ACCOUNT	ACCT
ACCOUNTS-RECEIVABLE	AR
ADDRESS	ADDR
AREA	AREA
BACKORDER	BORDR
BIN	BIN
BRANCH	BRNCH
BUDGET	BUDG
CALENDAR-YEAR	CALYR
CATEGORY	CATGY
CEILING	CEIL
CHANGE	CHNG
CHARGE	CHRG
CLASS	CLASS
COMMENT	COMNT
COMPANY	CO
CONTACT	CNTCT
COURSE	COUR
CREDIT	CRED
CUSTOMER	CUST
SALARY	SALRY

EXAMPLE

An attribute "Salary" has this description

"Salary paid to the employee from the beginning of the current year to present"

Attribute Name in full

Employee-Year-To-Date-Salary-Amount

Revised Attribute Name

EMP-YTD-SALRY-A

Document Attributes

- Name attribute using data naming standard
- What is the format?

  (e.g. Alphanumeric, numeric, decimal)
- How is it validated?

Attribute Analysis

ATTRIBUTE NAME	TYPE	LENGTH	DEFINITIONS AND BUSINESS RULES
EMP-M	A	25	Employee Name (full name, start with the surname)
Emp-JOB-T	A	25	Employee Job title (Programmer, Analyst, Project Manager, Department Manager)
EMP-JOB-DESC	A	60	Simple short description
PROJ-M	A	10	A unique short name given to the project
MTH-SAL-A	N	6.2	Monthly salary (999999.99)
EMP-PROJ-START-D	D	8	Start date of the project (DDMMYYYY)
EMP-PROJ-END-D	D	8	End date of the project (DDMMYYYY)

Note: A – Alphanumeric, N – Numeric, D - date

Uniqueness

- Sometimes it is necessary to introduce artificial attributes to
  - ensure uniqueness
    - example:
    - employee number as employee name may not be unique

Codes

- Sometimes it is necessary to introduce artificial attributes to
  - simplify identification
    - provide codes example country code, currency, etc
    - cut processing of large size attribute
    - Example:

COUNTRY CODE

SG	-	Singapore
CN	-	China
IN	-	India
MM	-	Myanmar
MY	-	Malaysia

Documentation

- Document the attributes in Data Dictionary

Attribute Assignment to Entities

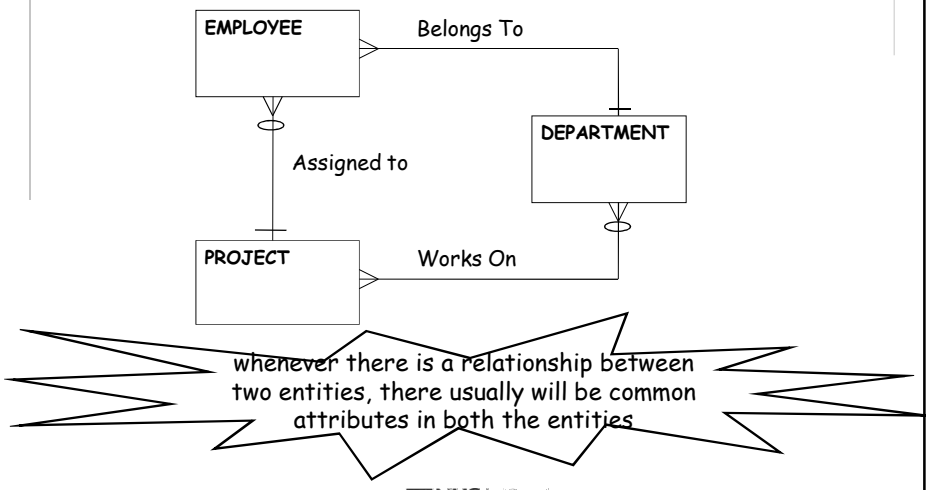
Attributes of the Project Management system

ENTITY	ATTRIBUTE NAME	TYPE	LEN	DEFINITION AND BUSINESS RULES
Employee	EMP-N	N	6	Employee number. Unique for each employee.
	EMP-M	A	25	Employee name in the form of the last name and two initials.
	EMP-JOB-T	A	25	Employee job Title (analyst, programmer, project manager, department manager)
	EMP-JOB-DESC	A	60	Short description of job title
	SAL-CHNG-D	N	8	The effective date of the employee's salary.. Employee salary history is kept for 3 years.
	MTH-SAL-A	N	6.2	Monthly salary (in the form of 999999.99).
	PROJ-M	A	10	Name of project currently assigned to employee Each project has a unique name
	EMP-PROJ-START-D	D	8	Start date of the employee on the project
	EMP-PROJ-END-D	D	8	End date of the employee on the project
	SKILL-TYPE-C	A	6	Skill type code
	SKILL-TYPE-DESC	A	20	Description of the skill (usually an employee has more than one skill

Attribute Assignment to Entities

ENTITY	ATTRIBUTE NAME	TYPE	LEN	DEFINITION AND BUSINESS RULES
Department	DEPT-M	A	10	A unique name for each department
	DEPT-MGR-N	N	6	Employee number of the manager in charge of department
	DEPT-MGR-M	A	25	Name of the manager in charge of department
	DEPT-EMP-SIZE-Q	N	3	Number of employee in the department
	DEPT-PROJ-M	A	10	Project (name) currently assigned to department
	DEPT-PROJ-BUDGT-A	N	6.2	Allocated budget(money) for the project for that department
	DEPT-EMP-N	N	6	Employee Numbers of all employees working in the department
	DEPT-EMP-M	A	25	Employee names of all employees working in the department
Project	PROJ-M	A	10	Unique name assigned to project. Project currently work on by company.
	PROJ-DESC	A	100	Short description of the project
	PROJ-BUDGT-A	N	6.2	Total budget (money) allocated to the entire project
	PROJ-START-D	D	8	Project start date (DDMMYYYY)
	PROJ-END-D	D	8	Project end date (DDMMYYY). The project must end by this date.

Attribute Assignment to Entities



Terminologies

	IBM	Codd	Traditional
Entity	Table	Relation	File
Entity Occurence	Row	Tuple	Record
Attribute	Column	Attribute	Field

Project

PROJ-N	PROJ-DESC	PROJ-BUDGT-A	PROJ-START-D	PROJ-END-D
Leave	In house leave application system	10000.00	01012003	01062003
FSTA Tour	Consultancy project on a Tour Booking System for FSTA	500000.00	01012002	01122002
AC System	Inhouse Accounting System	30000.00	01032002	01122003

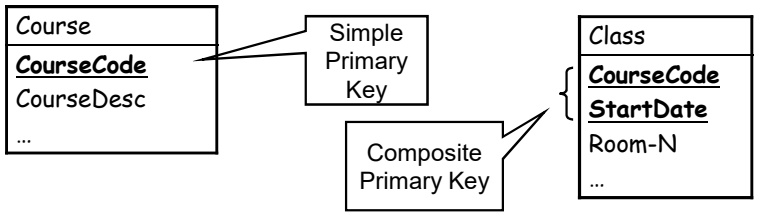
Identifiers/Keys

"...an entity **occurrence** must be **identifiable**"

- For each entity in the ERD, select a **primary key** from the list of attributes

Identifiers/Keys

- Primary Key:
  - A simple key is a key consisting of a single attribute
  - A composite key is a key consisting of more than one attribute



Identifiers/Keys

Restrictions

- No two rows must contain the same value for their keys
- None of the component attributes of the identifier may have null values.

Identify Key Attributes

ENTITY	ATTRIBUTE NAME	TYPE	LEN	DEFINITION AND BUSINESS RULES
Employee	EMP-N	N	6	Employee number. Unique for each employee.
	EMP-M	A	25	Employee name in the form of the last name and two initials.
	EMP-JOB-T	A	25	Employee job Title (analyst, programmer, project manager, department manager)
	EMP-JOB-DESC	A	60	Short description of job title
	SAL-CHNG-D	D	8	Date an employee's salary was changed, in the form of 'DDMMYYYY`. Employee salary history is kept for 3 years.
	MTH-SAL-A	N	6.2	Monthly salary after change on a given date (in the form of 999999.99).
	PROJ-M	A	10	Name of project assigned to employee Each project has a unique name
	EMP-PROJ-START-D	D	8	Start date of employee assignment to the project, DDMMYYYY (assignment period of employee to the project varies between employees)
	EMP-PROJ-END-D	D	8	End date of employee assignment to the project, DDMMYYYY (assignment period of employee to the project varies between employees)
	SKILL-TYPE-C	A	6	Skill type code
	SKILL-TYPE-DESC	A	20	Description of the skill

Identify Key Attributes

ENTITY	ATTRIBUTE NAME	TYPE	LEN	DEFINITION AND BUSINESS RULES
Department	DEPT-M	A	10	A unique name for each department
	DEPT-MGR-N	N	6	Employee number of the manager in charge of department
	DEPT-MGR-M	A	25	Name of the manager in charge of department
	DEPT-EMP-SIZE-Q	N	3	Number of employee in the department
	DEPT-PROJ-M	A	10	Project (name) assigned to department
	DEPT-PROJ-BUDGT-A	N	6.2	Allocated budget(money) for the project for that department
	DEPT-EMP-N	N	6	Employee Numbers of all employees working in the department
	DEPT-EMP-M	A	25	Employee names of all employees working in the department
Project	PROJ-M	A	10	Unique name assigned to project
	PROJ-DESC	A	100	Short description of the project
	PROJ-BUDGT-A	N	6.2	Total budget (money) allocated to the entire project
	PROJ-START-D	D	8	Project start date (DDMMYYYY)
	PROJ-END-D	D	8	Project end date (DDMMYYY). The project must end by this date.

Summary

- Data Analysis
  - Entity Relationship diagram (data model)
    - Entity
    - Relationship
  - Attribute Analysis
    - Identify and analyse attributes
    - Identify primary key
  - Documentation
    - Data Dictionary

Exercise – Purchase Order

Order Id: A1091

Purchase Order

Customer ID: S009
Customer Name: Lynn Wang
Date: 21/7/2011

ItemCode	Description	Qty
S1001	Pencil	100
S1003	Eraser	200
S1005	Ruler	250

Total Number of Items: 3

Product Code : S1001
Product Name: Pencil
Unit Price: $0.20

Business Rules:
- An order contain 1 to many products
- A product may be appear in multiple order
- A product may not be ordered if it is not popular!

Ex. Draw an ERD and identify attributes belong to the entities

last page